Computational Prediction and Design to Create Iteratively Larger Heterospecific Coiled Coil Sets

Richard O. Crooks, Alexander Lathbridge, Anna S. Panek, and Jody M. Mason*

[*]Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY

[1]To whom correspondence should be addressed: j.mason@bath.ac.uk

Running title: Designing heterospecific Coiled coils

Keywords: protein-protein interactions, *de novo* design, heterospecific proteins, interactome screen, computational biology, library screening, coiled coils

A major biochemical goal is the ability to mimic nature in engineering highly specific protein-protein interactions. We previously devised a computational interactome screen to identify eight peptides that form four heterospecific dimers despite 32 potential off-targets. To expand the speed and utility of our approach and the PPI toolkit, we have developed new software to derive much larger heterospecific sets (≥ 24 peptides) while directing against antiparallel off-targets. It works by predicting T_m values for every dimer based on core, electrostatic, and helical propensity components. These guide interaction specificity, allowing heterospecific coiled coil sets to be incrementally assembled. Prediction accuracy is experimentally validated using circular dichroism and size exclusion chromatography. Thermal denaturation data from a 22 coiled coil (CC) training-set was used to improve software prediction accuracy, and verified using a 136 CC test-set consisting of 8 predicted heterospecific dimers and 128 off-targets. The resulting software, qCIPA, individually now weighs core *a-a'* (II/NN/NI) and electrostatic *g-e'*$^{+1}$ (EE/EK/KK) components. The expanded dataset has resulted in emerging sequence context rules for otherwise energetically equivalent CCs; for example, introducing intra-helical electrostatic charge-blocks generated increased stability for designed CCs while concomitantly decreasing the stability of off-target CCs. Coupled with increased prediction accuracy and speed, the approach can be applied to a wide range of downstream chemical and synthetic biology applications, in addition to more generally to impose specificity in structurally unrelated PPIs.

Introduction

Protein structures and their interactions form via complex arrangements of cooperative interactions, making *de novo* design of heterospecific Protein-Protein Interactions (PPIs) very difficult. There is a large shortage in the number PPI components that are increasingly needed in biological applications, where specificity of interaction is important and where large numbers of heterospecific peptide pairs would be of benefit[1, 2]. For example, this unmet need in protein science includes applications in peptide labelling (e.g. monitoring biochemical processes without the need for large tags such as GFP); in delivery of drugs or toxins; in protein purification and labelling applications as high specificity affinity-tags; in creation of large nanostructures such as tetrahedral cages or conductive nanowires; in biomaterials such as reversible hydrogels that assemble or disassemble according to pH or temperature change; in disease modulation, and many other uses as specific cognate pairs in the synthetic biology toolkit [3-5]. Specificity of protein-protein interaction and recognition is also essential for normal physiology, with protein-interaction network imbalances associated with a wide range of diseases. A major drawback in applying proteins and peptides to such applications is the limited number of exquisitely specific orthogonal PPI forming peptides that are available. This is because, despite considerable effort, sequence to structure information relating to the protein-folding problem is largely unsolved [6]. However, this is becoming possible for systems such as coiled coils (CCs), where the rules translating how primary sequence dictates quaternary structure are becoming increasingly understood [7, 8]. The CC motif is an interesting PPI model as it is a simplistic example of quaternary structure and commonly found in a wide range of therapeutically relevant proteins. Utilising CCs as a model to predict the stability and specificity of protein dimerisation directly from the primary sequence is therefore an important and tractable goal. This is because despite apparent simplicity, CCs are highly specific in the interactions that they drive. Using knowledge of this type of protein fold we have used *de novo* design to generate the formation of specific CCs that can be applied in a wide range of applications. Here we utilise newly created software to allow great expansion in the number of specific CC forming peptides, and produce large customised sets of peptides that vary according to the users needs. To meet these aims we have built and tested freely available computational tools (see Supporting Information) that allow the user to derive large numbers of structurally similar orthogonal pairs with the potential to create excellent candidates for scaffold parts.

Designing coiled coil pairs. Although a good qualitative understanding exists for sequences that form a parallel dimeric CC, a quantitative understanding of how precise residue placements within dictate both stability *and* specificity is still lacking. We use a combination of known free energies derived via double mutant analyses for electrostatic ***g-e'***$^{+1}$ interactions [9] and predominantly hydrophobic ***a-a'*** interactions [10, 11], and combine these with general amino acid properties such as helical propensity [12],

to predict the T_m of a parallel dimeric CC given only the sequences of the constituent peptides. Optimising these parameters using our growing training set of experimentally tested CCs has allowed us to refine our software to make more accurate predictions that can then be tested on a much larger data set derived using protein arrays [13]. These new bioinformatics tools for CC prediction began with the *bZIP coiled coil interaction prediction algorithm (bCIPA)* [14, 15]. bCIPA was derived to estimate the T_m of a given parallel dimeric CC using ***only*** the primary sequence and was shown to correctly predict 97% of all strong interactions and 95% of all non-interacting pairs using an independent data set of human bZIP proteins [13]. This prediction was more accurate than a previously published prediction program [16] and utilized very simple and easily adaptable scoring matrices. Unlike related qualitative algorithms [13 16], bCIPA makes a quantitative estimate by predicting a T_m value for an interaction between two component polypeptide chains. The approach is distinct from more recent work by the Keating group [1, 17-19], which make predictions via complex computational algorithms using integer linear programming and cluster expansion to generate peptide ligands for defined targets / off-targets. Although this software does not consider antiparallel dimers, the group have created bespoke software that does[20]. Similarly the Woolfson group derived the CCBuilder software to study CCs by generating backbones, building in side chains, and providing atomistic models and a range of metrics on which to test their designs [21]. In contrast to the above our software is web-based, user-accessible, and differs in that it searches within large user-defined peptide sets to identify and provide quantitative outputs in the form of a T_m to derive heterospecific CC sets while directing against antiparallel CC alignments.

Building on our work in this area [22], we are constructing and expanding a suite of software to meet the goal of identifying very large sets of orthogonal pairs starting from large peptide libraries (Figure 1). In turn these results are being used to further refine accuracy of prediction while facilitating new biological understanding and an expansion in use by the wider scientific community. These user-friendly tools complement experimental work, and will allow for the development of designed CC motifs that are highly specific and that have a wide-range of potential downstream applications alluded to above. By experimentally testing *in silico* predictions, we demonstrate effectiveness in providing new and expanded heterospecific sets, while concomitantly refining the software for the design and creation of customised sets that vary in stability according to the needs of the user.

Materials and Methods

Design Rationale **–** The peptide library contained semi-randomised residues at all four ***a***, ***e***, and ***g*** positions within the heptad repeat of the 37mers[22]. Options of Glu and Lys were included at all ***e*** and ***g*** positions. Lys was used since it has comparable performance to Arg in terms of helicity and forming electrostatic interactions, but is easier to incorporate into synthetic peptides. Gln, used in previous libraries and designs [14, 23] is omitted since it interacts favourably with both acidic and basic residues

and is not therefore expected to confer significant specificity to pairings, and would therefore be expected to be selected out during the screening. At ***d*** positions, Leu was maintained throughout as these are known to assist in driving the formation of parallel and dimeric CC species. At ***a*** positions, the residues were semi-randomised to Asn and Ile. These residues provide the greatest specificity distinction between core position residues based on double mutant analyses [11], with Asn-Asn (-2.4 kcal/mol) and Ile-Ile pairs (-9.2 kcal/mol) both significantly more favourable than an Asn-Ile pair (-0.5 kcal/mol). These energetic values are anticipated to give a specificity enhancement caused by favourable alignment relative to misaligned residues. Therefore, Asn-Asn pairing confers specificity because the hydrogen bonding benefit outweighs the lack of stability and limits oligomeric states to dimers [24, 25]. Asn-Asn and also Ile-Ile ***a-a'*** pairs are predicted to stabilise the derived peptides as dimers rather than higher order oligomers or antiparallel CCs, where Asn-Asn core pairings are also not found [26]. This is because ***a-a'*** and ***d-d'*** contacts occur in parallel but not antiparallel CCs, meaning that an interaction between equivalent Asn residues in a homodimer will favour a parallel alignment [27]. Furthermore, it is anticipated that alignment of Asn residues in core positions will stabilise a particular axial alignment, and prevent alternative axial alignments causing unexpected interaction patterns.

In Silico Library Screening – The *in silico* library was created using ***Generate Library Sequences*** (see SI) to list each user-defined member of the library in a sequential manner. This library was next screened using the ***bCIPA interactome screen*** engine (see SI), which was developed to screen interactomes of sequences using bCIPA [14, 15] and derive a heatmap for millions of hypothetical peptide pairs. A 4,096 member peptide interactome was reduced to 1,536 by specifying that a minimum of two Asn and two Ile residues are required at ***a*** positions to assist in imposing specificity. The resulting 1,180,416 hypothetical pairwise interactions within it were next screened using ***Find pairs*** (see SI) to identify groups of four sequences which when placed together would be predicted to form heterospecific dimeric interactions, known as 'pairs'. These pairs could then be further screened within the same page (using ***Find Quadruples***) to identify groups of eight sequences which when placed together in solution would again be predicted to form four heterospecific dimeric interactions, known as 'Quadruples'. Finally, Quadruple sets were combined to identify sets of sixteen peptides able to form eight heterospecific CCs (***Find Octuples***).

Sequence Screening Protocol – Sequences which met the conditions of the initial constraints described were retained for the interactome screen. These were specificity against homodimerisation and the requirement of two Asn residues. The latter has been used previously to create heterospecific sets[22, 28] since it maximises the potential for specificity in desired pairs where core NI pairings are energetically much less favoured that NN or II [11]. Elimination of sequences which do not fulfil these requirements at the outset reduces computational load, allowing even larger libraries to be screened than in the presented example. Each new sequence that satisfied these criteria was added to the array

and screened using the *bCIPA interactome screen* engine for interaction affinity with every other sequence in the array. This occurred at the time the sequence was added and prevents any repeated calculations so that each interaction is only calculated once (i.e. not bi-directionally). The results of these calculations were stored in the database, but only if the affinity of those interactions exceeded the minimum specified affinity of the desired heterospecific pairs (in this case 70 °C). Thus this database was a list of pairs of sequences, which could potentially form heterospecific pairings. Interactions in this database, with a T_m greater than the minimum allowed in the input were paired with each other iteratively, with a computational load-saving requirement that excluded pairs from being screened against one another where those pairs contained any of the same peptides (e.g. an interaction between peptides 1 and 2 could not be paired with an interaction between peptides 1 and 3, since peptide 1 appears in both interactions such that the pairs would not be specific, as there is clear cross talk without needing to quantify the interactions). Potential pairs which did not have any identical sequences were paired iteratively, in a similar manner to identifying the peptide pairs. However instead of a simple bCIPA calculation, a mini-interactome was created for each potential pair and the T_m calculations of interactions contained therein were checked against a user specified maximum undesired T_m. Any undesired interactions with a predicted T_m of greater than 20 °C meant that the group of sequences was rejected as a specific pair. Where sequences met these criteria, they were retained as a pair of non-interacting CCs identified in the interactome. Quadruples were next identified by comparing sets of pairs to one another in a similar manner as previously (by cross-checking identified non-interacting pairs). However, in the case of Quadruples, the increased stringency meant that a higher maximum T_m for an undesired interaction was used, in this case 30 °C, with a minimum ΔT_m (desired – non-desired) of 40 °C.

Screening Parameters – To generate sets of 16 peptides predicted to form eight heterospecific CCs, the maximum acceptable predicted T_m for homodimers was set as 10 °C (this value dictates the number of non-homodimeric peptides permitted to proceed into the main screen), the minimum T_m for desired heterodimers as 70 °C, the maximum T_m for undesired heterodimers as 20 °C, and the minimum ΔT_m (desired – off-target) as 50 °C. Further increasing stringency resulted in fewer initial peptides that progressed to Octuples, or resulted in many lower stringency sets (i.e. lower ΔT_m (desired – off-target)) that therefore took significantly longer to identify. These parameters resulted in the software identifying 42 separate pairs of predicted non-interacting CCs. The highest predicted T_m for desired CCs was 73 °C and the highest predicted T_m for undesired CCs was 18 °C. Having identified two heterospecific CCs, the program combined pairs to identify 72 sets of four CCs (Quadruples). Next, a minimum ΔT_m of 21 °C, and a maximum undesired CC T_m of 52 °C was specified within the software. This resulted in the retention of 72 groups of non-interacting Quadruples with a lowest desired T_m of 73 °C and a highest undesired CC T_m of 28 °C. Finally the same parameters were used to combine quadruples in identifying eight CCs (Octuples). This resulted

in 36 groups of non-interacting Quadruples of CCs, with a lowest desired T_m of 73 °C and a highest undesired T_m of 52 °C.

Homodimer removal – In order to preserve system resources and to limit the interactome screen to within useful search space, sequences which were not expected to produce specific CCs were removed. Search constraints for the interactome excluded all sequences which were predicted to have a homodimeric T_m greater than 10 °C at the earliest opportunity (as sequences are imported into the script). Sequences retained at this stage were stored in a MySQL database, together with the Williams helicity score [12] (to save recalculation).

Antiparallel CC removal – We have enabled a new feature that searches for and removes homodimers that generate full electrostatic complementarity in the antiparallel orientation. We previously noted that antiparallel dimers were not predicted to form owing to that fact that Asn-Asn core pairings between ***a-a'*** residues that make the major energetic contribution to CC specificity in the parallel orientation are unable to do so in the antiparallel orientation[22]. Rather, buried polar interactions in antiparallel dimers take place between ***a-d'*** residues and would therefore not be considered possible in this system [29, 30]. This approach has been used previously to direct against antiparallel dimer formation for heterospecific sets [28]. However, we previously speculated that this was not enough to direct against potential antiparallel orientations that result in fully complementary electrostatics (i.e. e-e' or g-g') [22]. Directing against full electrostatic complementarity in the antiparallel orientation therefore provides an additional barrier to removing these otherwise permissible antiparallel pairs. It also reduces the search time of the algorithm by increasing the stringency in the selection of the initial sequences that are processed into interactions, and consequently reduces the size of the search required to find pairs and Quadruples.

Peptide synthesis – Rink amide ChemMatrix™ resin was obtained from PCAS Biomatrix, Inc. (St.-Jean-sur-Richelieu, Canada); Fmoc-L-amino acids and 2-(1H-Benzotriazole-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate (HBTU) or benzotriazol-1-yl-oxytripyrrolidinophosphonium hexafluorophosphate (PyBOP) were obtained from AGTC Bioproducts (Hessle, UK); all other reagents were of peptide synthesis grade and obtained from Thermo Fisher Scientific (Loughborough, UK). Peptides were synthesized on a 0.1-mmol scale on a PCAS ChemMatrix™ Rink amide resin using a Liberty Blue™ microwave peptide synthesizer (CEM; Matthews, NC) employing Fmoc solid-phase techniques (for review see [31]) with repeated steps of coupling, deprotection and washing (4 × 5 ml dimethylformamide). Coupling was performed as follows: Fmoc amino acid (5 eq), HBTU OR PyBOP (4.5 eq), and diisopropylethylamine (10 eq) in dimethylformamide (5 ml) for 5 min with 35-watt microwave irradiation at 90 °C. Deprotection was performed as follows: 20% piperidine in dimethylformamide for 5 min with 30-watt microwave irradiation at 80 °C. Following synthesis, the peptide was acetylated – acetic anhydride (3 eq) and

diisopropylethylamine (4.5 eq) in dimethylformamide (2.63 ml) for 20 min – and then cleaved from the resin with concomitant removal of side chain-protecting groups by treatment with a cleavage mixture (10 ml) consisting of TFA (95%), triisopropylsilane (2.5%), and H_2O (2.5%) for 4 h at room temperature. Suspended resin was removed by filtration, and the peptide was precipitated using three rounds of crashing in ice-cold diethyl ether, vortexing and centrifuging. The pellet was then dissolved in 1:1 MeCN/H_2O and freeze-dried. Purification was performed by RP-HPLC using a Phenomenex Jupiter Proteo (C18) reverse phase column (4 μm, 90 Å, 10 mm inner diameter × 250 mm long). Eluents used were as follows: 0.1% TFA in H_2O (A) and 0.1% TFA in MeCN (B). The peptide was eluted by applying a linear gradient (at 3 ml/min) of 5% to 70% B over 40 min. Fractions collected were examined by electrospray mass spectrometry, and those found to contain exclusively the desired product were pooled and lyophilized. Analysis of the purified final product by RP-HPLC indicated a purity of >95%.

Circular Dichroism – CD was carried out using an Applied Photophysics Chirascan CD apparatus (Leatherhead, U.K.) using a 200 μl sample in a CD cell with a 1 mm path length. Samples contained 150 μM total peptide (Pt) concentration at equimolar concentration for heterodimeric solutions (i.e. 75 μM per peptide) and suspended in 10 mM potassium phosphate and 100 mM potassium fluoride at pH 7 one hour prior to analysis. The buffer was chosen to be CD-compatible while being close to physiological pH and salt conditions. The CD spectra of samples were scanned between 300 nm and 190 nm in 1 nm steps, averaging 0.5 s at each wavelength. Three scans at 20 °C were averaged to assess helical levels and CC structure.

Thermal Denaturation Experiments – Thermal denaturations were performed at 150 μM P_t in 10 mM potassium phosphate, 100 mM potassium fluoride, pH 7, using an Applied Photophysics Chirascan CD instrument (Leatherhead, U.K.). The temperature ramp was set to stepping mode using 1ºC increments and paused for 30 seconds at each temperature before measuring ellipticity at 222 nm. For all temperature denaturation experiments data collection was started at -8 °C, and at this temperature, the peptide solutions remained aqueous. Data collection continued to 95 °C. Data points for thermal denaturation profiles represent the averaged signal after 4 s of data collection. Melting profiles (Figure 2) were ≥95% reversible with equilibrium denaturation curves fitted to a two-state model, derived via modification of the Gibbs-Helmholtz equation [14, 32, 33], to yield the melting temperature (T_m). Melting profiles for heterodimers are clearly distinct from averages of constituent homodimeric melts (Figure 2 and 3), indicating that helices form heterodimeric complexes, with the cooperative nature of the melting profiles suggesting an apparent two-state process. T_m values were determined by least-squares fitting of the denaturation assuming a two-state folding model that is widely used for CCs [33] and provided an excellent fit to our data.

Size Exclusion Chromatography – Size exclusion experiments were performed at room temperature using a Superdex Peptide 10/300 GL column (GE Healthcare Life Sciences) by injecting 100 µl of a 50 µM or 10 µM (total peptide concentration) sample in 10 mM potassium phosphate, 100 mM potassium fluoride, pH 7 at a flow rate of 0.5 ml/min. Elution profiles were recorded via A_{280}.

Results and Discussion

We previously generated a 1,536 member computational library of peptides 32 residues in length, and successfully screened using a PHP-based algorithm it to predict the formation of four heterospecific parallel dimeric CCs[22]. Here we describe screening the resulting 1,180,416 CC interactome ([1536*1537]/2) using a much faster and more expansive Python-based algorithm that has enabled the identification of many different sets of sixteen peptides that when combined are predicted to form eight heterospecific dimeric CCs. In doubling the number of desired heterospecific pairs from four to eight, the number of off-targets are quadrupled from 32 ([8*9/2]-4) to 128 ([16*17/2]-8), making this a particularly challenging task (Figure 1). The algorithm is further improved over the previous version in that it removes peptides predicted to form antiparallel CCs. Dimerisation is driven by Lys/Glu options at every ***e*** and ***g*** position and Ile/Asn options at every ***a*** position within the heptad repeat, creating the necessary options to direct the formation of heterospecific CC sets [10, 11, 34]. The ***d*** positions were fixed as Leu to further direct formation of parallel and dimeric CCs[24, 35], with remaining positions fixed as Ala to promote α-helicity. Screening works by iteratively identifying within a set of sequences which CCs are and are not predicted to form using a set of temperature cut-offs input by the user. The program assigns a predicted T_m for every hypothetical CC within the interactome and creates an associated heat-map. Stringency of screening can be directed by inputting the required T_m for desired pairs, as well as the T_m cut-off for homodimeric and heterodimeric off-targets (see Methods and SI for more details).

Combining Core and Electrostatic arrangements to confer stability and specificity. The number of energetic arrangements in the Octuples (eight CC) set is expanded from those previously observed (Table 1). Setting the desired CC cut-off T_m as high as possible (70°C) led the software to arrive at fully complementary electrostatic arrangements and fully optimal 2xII/2xNN core arrangements. For homodimeric off-targets, the same core arrangements were observed as for desired pairs, but with fully repulsive electrostatic (***g-e'***$^{+1}$) arrangements. For intra-pair off-targets (i.e. within a designed interactome of two CCs), the core consisted of a 4NI/IN (fully mismatched) arrangement with four attractive (KE) and four repulsive (EE/KK) electrostatic interactions. However, the inter-pair off-target arrangements (interactions outside of designed interactomes of two CCs) were expanded in the Octuples set relative to our previously observed combinations [22]. In the Quadruples sets (i.e. four CCs;

peptides 1-8 or peptides 9-16), we observe a 1II/1NN/2NI core arrangement with the same four attractive (KE) and four repulsive (EE/KK) electrostatic arrangement. This was the most commonly observed off-target configuration, accounting for 50% of all off-targets. In combining Quadruples to arrive at Octuples, three additional scenarios were observed (i.e. in CC interactions between peptides from set 1-8 with set 9-16). These additional off-target combinations were ***i)*** an optimised core (2NN 2II) with 4 attractive and 4 repulsive electrostatic arrangements, leading to a further increase in the predicted stabilities of off-targets. ***ii)*** a fully mismatched 4NI core arrangement with an all-attractive electrostatic arrangement. ***iii)*** a fully mismatched 4NI core arrangement with an all-repulsive electrostatic arrangement. The last two scenarios were predicted by bCIPA as being either very stable (52 °C) or very unstable (-29/-21 °C), which we believed to be over and under estimated respectively. Therefore a set of heterospecific Octuples could be identified, albeit with lower overall stringency than for Quadruples owing to the increased stability off-targets highlighted above (Table 1 and Figure 3). The data from an exhaustive low-complexity set of options (i.e. all Asn/Ile core options and all Glu/Lys electrostatic options) is therefore sufficient to generate the required number of unique core/electrostatic arrangements for the creation of such larger heterospecific sets. Moreover, the large amount of data generated in predicting and experimentally testing an Octuple heterospecific set has allowed creation of a more refined version of software. Subsequently, this can be used to predict heterospecific interactomes for this particular subset of CC. This would strengthen prediction of affinity for individual pairs and therefore allow the creation of both larger and more accurate heterospecific sets.

Experimental Characterisation of Coiled coils. To demonstrate that *in silico* generated sequences are specific *in vitro*, sixteen peptides predicted to form eight heterospecific CCs were synthesised and characterised. CD spectra and associated thermal denaturation experiments were used to establish that all samples displayed characteristic α-helical profiles and to determine the T_m value for each CC within the sixteen peptide interactome, and therefore the relationship between predicted and measured values (Figure 3). In these experiments, seven of eight CCs predicted to be heterospecific were verified experimentally, with predicted T_m values of 73 °C found to be accurate to within 5°C, with the exception of CC 15-16, which was 21 °C lower than expected. The 128 off-target interactions had predicted T_m values of -8 °C to 52 °C, which were measured to range from -3 °C to 60 °C, ensuring that the 1-2, 3-4, 5-6, 7-8, 9-10, 11-12, and 13-14 interactome was heterospecific as designed, with all 98 off-targets disfavoured. As can be observed, differences in molar ellipticity at pre-melt temperatures reflect the fact that helicity is only one determinant in CC stability; side-chain preferences at core and electrostatic positions are other key determinants that bCIPA uses when determining the Tm.

As a further demonstration of correct peptide pairing size exclusion chromatography was used (SEC; Table s8). As shown previously [22], monomeric elution profiles are superimposable, and occur

at 20 minutes for all homomeric solutions, indicating that all 16 peptides are monomeric at room temperature and a total peptide concentration of 50 μM. In contrast, at 50 μM the profiles for all 8 desired heterodimers eluted at 19 minutes. In both cases the elution profiles were consistent with predicted monomer/dimer patterns. In addition a number of off-target samples were run, demonstrating that those with a measured T_m <50 °C were monomeric at 50 μM, whereas those with a T_m >50 °C ran as a dimer at 50 μM total peptide concentration. However, at a concentration of 10 μM, all desired peptides (T_m = 68-75 °C), with the exception of 15-16 (T_m = 52 °C), were found to remain in complex as dimers. At the same concentration all nine off-target samples with a measured T_m>50 °C (i.e. T_m = 51-60 °C) were found to have shifted to monomeric samples. As SEC controls two peptides of similar length were selected that have been previously characterised and shown to exist in either monomer form (20 mins), or as a parallel dimeric CC (19 mins) [14, 22].

Comparison of Predicted and Observed data – bCIPA. Throughout the process T_m values for predicted and observed pairs were compared to improve the accuracy of prediction. When comparing CCs selected together in pairs, where interactomes consist of ten potential CCs (e.g. peptide sets 1-4, 5-8, 9-12, or 13-16) there was generally an excellent correlation. This was the case in both 22 previously characterised CCs (overall r^2 = 0.6 [22]; r^2 = 0.69 for peptide set 1-4 within) as well as when the fit was applied to newly derived peptides sets 1-4, 5-8, 9-12, and 13-16 within the 136 CCs. The strong correlations (r^2 = 0.70 to 0.82) observed relates to the fact that desired pairs have very high T_m values (e.g. two CCs of approx. 73 °C), and four homodimeric off-targets have very low T_m values (four potential CCs of approx. 0 °C). The remaining four members of each *in silico* selected four-peptide interactome display predicted T_m values in the 20-30 °C range. Therefore each resulting ten CC interactome contains a wide T_m range to which the subsequent fit is strong. Fitting to the off-targets more generally (particularly inter-pair off-targets) is more challenging since the predicted ΔT_m is very narrow (e.g. typically just 10°C for the majority of inter-pair off-targets). This means that similar variations in the predicted temperatures yields much lower r^2 values. However, importantly, the predicted *vs*. observed T_m values for each type of interaction demonstrates that the general goal of heterospecificity is achieved (Figure 3).

Issues with the Computational Approach. Despite successfully demonstrating that PHP-based bCIPA software can predict many hypothetical sets of Octuples, we experienced several issues that limited its further implementation. These were speed; it was time-consuming for larger peptide sets and not expansive enough to identify larger numbers of these sets by relaxing the screening parameters. In addition, antiparallel options were not removed. Finally, the accuracy of the bCIPA prediction algorithm left room for improvement in these low sequence diversity peptides. These problems have been partially dealt with by moving away from PHP-based software to Python-based architecture. Previously, results were uploaded and processed on an external web-server causing the program to become slow as the number of peptides increased, limiting the number that could be progressed

within the search. Using a Python-based system, outputs are now simply written as text files and saved locally by the user, saving computational resource in the process. For instance, increasing the permitted homodimeric T_m by only a few degrees substantially increases the size of the interactome to be searched. The Python-based approach has enabled these larger data sets to be processed in a much shorter space of time. The initial check now removes potential antiparallel homodimers by removing those that result in an all-attractive electrostatic component. This additionally prevents potential heterodimeric antiparallel CCs from entering the interactome search. Removing potential parallel homodimers and antiparallel homo/heterodimers initially, restricts the number of peptides that enter the main interactome search, further reducing the redundancy of the system.

Refining bCIPA to improve T_m prediction for specific residue pairings – qCIPA. We have previously shown that bCIPA can accurately predict the thermal stability of CC pairs that are diverse in sequence [14, 15] [14] and that it can be used to generate *in silico* interactome predictions to guide the derivation of heterospecific CC sets[22]. The utility of bCIPA was demonstrated using a small eight-peptide interactome to derive four parallel dimeric CCs that were heterospecific when combined, despite 32 off-target CCs that could potentially associate. Using a completely new set of peptides we expanded this approach to a 16-peptide interactome. In doubling the number of desired heterospecific CCs from 4 to 8, the number of off-targets quadruple from 32 to 128, leading to a significant increase in the complexity of the design process. In turn, as the number of attractive and repulsive permutations becomes exhausted, higher stability off-targets must be included in the interactome (see above), leading to a decrease in the stringency of the system. Thus predicting larger heterospecific sets is a challenging task that requires a high accuracy of prediction. This is because any decrease in stringency will increase the likelihood of identifying off-targets of similar stability to the desired pairs. We have been largely successful in these aims and have arrived at an interactome of fourteen peptides that form seven heterospecific CCs despite 98 potential CC off-targets. In particular, CC 15-16 was found to display a lower T_m that is close to some of the off-target interactions. The results from this study further highlight the strengths and weaknesses of bCIPA. We have used our previous interactome dataset of 8 independent peptides as well the 16 newly predicted peptides presented here to facilitate the creation of new customised software, known as qCIPA. qCIPA predicts the T_m with greater accuracy than bCIPA for the subset of CCs we describe here and which are closely related in sequence (Figure 3). qCIPA was devised using 22 previously characterised CCs as a training-set (r^2 = 0.69;). The resulting fit was then applied to the 136 CC test-set resulting from the 16 peptide interactome described here - with no sequence repetition between the two sets. The resulting correlation coefficients therefore provide a direct comparison with bCIPA (Table s9). On average, qCIPA provides a 3 °C improvement in prediction using the 136 CC test-set. In reducing the test-set interactome into its substituent interactions, there is a 9 °C improvement in predicting the 16 homodimers, a 4 °C improvement in predicting the 16 intra-pair off targets and a 3 °C improvement in

predicting the 96 inter-pair off-targets. This comes at a cost of a 3 °C deteoration in predicting the heterospecific pair interactions (Figure 3). Considering the end-goal of creating heterospecific sets, in which off-target interactions should be carefully avoided, this overall increase in accuracy is a welcome step forward. qCIPA works in a similar way to bCIPA in that the T_m is calculated as a function of core pairings (C), electrostatic pairings (ES) and helical propensity (HP). For bCIPA:

$T_m = (a*\Sigma HP) + (b*C) + (c*ES) + d$ (Eq 1)

Where a=81.33, b=-10.18, c=-4.78, and d=-29.13[15]. bCIPA considers a wide range of residues at a/d/e/g positions (LINVRKT/LINVRKT/KRDEQNALT/KRDEQNALT). In contrast for qCIPA the options at these positions are limited to IN/L/EK/EK. This allows the six exact pairings at **a-a'** and ***g-e'***$^{+1}$ to be explicitly described. Therefore in the equation used to determine T_m the core and electrostatic components are expanded so that each interaction has its own coefficient. For qCIPA:

$T_m = (a*\Sigma HP) + (b*II) + (c*IN) + (d*NN) + (e*EE) + (f*EK) + (g*KK) + h$ (Eq 2)

Where a=4.16, b=-1.75, c=11.78, d=-5.24, e=-11.30, f=-0.97, g=-76.22, h=30.18. As observed in Figure 3, these changes result in an improved fit to the 22 CC training set (i.e. the best fit that is used to derive values of a-h; overall $r^2 = 0.69$). Having obtained the above values, the consequent fits to the CC pairs in the 136 CC test-set could be obtained ($r^2 = 0.89$, 0.69, and 0.84 for pairs 1-4, 5-8, and 9-12, and 0.59 for 13-16).

As the data set continues to grow we predict that it will be possible to take further parameters into account, such as sequence specific context where the core and electrostatic contributions are equivalent but their positioning within each helix leads to increased or decreased stability above or below what is otherwise predicted. We have seen this previously for positive or negative residues at the helix termini that serve to stabilise or destabilise the helix macrodipole leading to over or underestimated stability [22, 36, 37]. At present there is insufficient data to build these predictions into our models, although general patterns within the data are emerging (see below).

Comparing Old and New Approaches. To generate Octuples (including the set presented here), the PHP-based bCIPA software was used to generate two heterospecific CCs and consequently Quadruples that were predicted to remain heterospecific when all component peptides are combined. This resulted in 72 unique sets of two CCs (setting minimum delta as 50°C and maximum off-targets permitted to progress into the interactome search as 20°C) and 144 sets of Quadruples (setting minimum delta as 40°C and maximum off-targets as 30°C). By repeating these steps while removing potential antiparallel pairs the numbers decreased to 42 unique sets of two CCs and 72 unique sets of Quadruples. To continue using this approach, unique sets of Quadruples were screened against each other to identify unique sets of Octuples. However, this was time consuming and led to only 36 hypothetical Octuple sets.

Using the faster Python-based qCIPA software now proposed, the algorithm creating heterospecific sets by screening one CC against another, one CC at a time (e.g. One CC→ two → three → four (quadruples) → five → six → seven → eight (octuples)), until no further unique sets can be identified. For example, the 1,536 peptides scanned for heterospecific pairs took 6 seconds using qCIPA in Python. In contrast bCIPA benchmarked at 42 seconds on the same machine when using PHP. This has led to many more unique sets being identified because the stringency in taking smaller increments is much lower and therefore more peptides are permitted to progress at each step. For example, using slightly less stringent parameters (the maximum T_m for undesired heterodimers and ΔT_m (desired – off-target); see Table 2), the 510 unique sets of pairs created 15171 unique sets of Quadruples. This 210-fold increase in the number of unique sets created 27501 unique sets of Octuples, over 760 fold more than previously identified. Coupled with this iterative approach, relaxing the homodimer stringency from the very start of the search procedure (Table 2) can be continued. Although not tested experimentally, this iterative process we have taken this as far as 54 unique sets of predicted duodecuples (12's), where none could be identified using our previous PHP approach [22]. It is important to note that while every set is unique, there are many instances of the same peptide occurring within multiple sets. This apparent redundancy in the search procedure is however necessary to ensure that sequences are retained during each iteration and that the highest possible number of heterospecific CC sets can be identified going forward. Relaxing the stringency further will increase this number until all core/electrostatic arrangements have been saturated, while significantly lengthening the search time from several days (e.g. 4 days in the case above) to many weeks on a standard PC.

Sequence Specific Context – Core. If we ignore sequence context then many pairs appear energetically identical in terms of core and electrostatic contributions (using helix propensity [12] and Core [10, 11] and electrostatic [9] scores calculated by Vinson and co-workers). The absence of sequence context calculations is reflected in the lack of diversity in predicted T_m values. While currently difficult to build into a qCIPA feature explicitly, we observe some general rules relating to sequence context that can be taken into account in future design rounds. This could be achieved, for example, by allowing the software to search libraries that conform to these rules in the first instance, such that the sequence specific peptides no longer need to be explicitly 'searched for'. In analysing the data, by grouping sequences with identical electrostatic arrangements, we are able to make some limited interpretations regarding the effect of core arrangements:

- An NN II NN II arrangement appears to lead to more stability than an II NN II NN arrangement. For example O5-6 > O15-16 (ΔT_m = 19), O13-14 > O7-8 (ΔT_m = 7), O5-5 > O15-15 (ΔT_m = 5), O14-14 > O8-8 (ΔT_m = 22), O6-6 > O16-16 (ΔT_m = 23), and O13-13 > O7-7 (ΔT_m = 22).

- In contrast to this, inspection of the data suggests that NN NN II II and II II NN NN are energetically equivalent. For example, O9-10 ~ O3-4 and O1-2 ~ O11-12.

However, both of these arrangements are predicted to stabilise desired states and off-targets by an equal amount, meaning that there is no preferential core arrangement in maximising ΔT_m values and therefore in achieving heterospecific CCs.

Sequence Specific Context – Electrostatics. As there are many examples of alternative electrostatic arrangements with identical core arrangement (Tables s1-7), we are able to make some general observations. When normalising for identical cores it becomes apparent that:

- For desired pairs (Table S1), blocks of same charge on either ***e*** or ***g*** residues of each peptide led to increased stability over that purely based on the sum of the core and electrostatic components [9-11]. For example, O13-14 > O5-6 (ΔT_m = 4) and O7-8> O15-16 (ΔT_m = 16) which together suggest that the electrostatic ***g-e'***$^{+1}$ KE KE KE EK arrangement is more stable than EK EK KE EK. Similarly O1-2 > O9-10 (ΔT_m = 3) and O11-12 > O3-4 (ΔT_m = 5), both suggesting that KE EK EK EK is more stable than EK KE EK EK. Taking this further, both O1-2 and O9-10 are favoured over Q1-2 (ΔT_m = 9 and 6) suggesting that EK KE KE EK is less stable than either KE EK EK EK or EK KE EK EK arrangements. Collectively this suggests that ***for desired pairs, intra-molecular repulsion between heptads increases intermolecular attraction and therefore increases CC stability***. We speculate that these intra-molecular 'charge blocks' at ***e*** and ***g*** positions within component helices increase stability for these desired pairs by promoting inter-molecular attraction. This may be due to the fact that intra-molecular repulsion between ***e-e***$^{+1}$ or ***g-g***$^{+1}$ residues helps to increase inter-molecular attraction between ***g-e'***$^{+1}$ pairs.
- Similarly, for intra-pair off-targets (Table S3), placing opposing inter-molecular charge repulsions next to each other (i.e. E followed by K at consecutive ***e*** or ***g*** positions within the same peptide) is more stabilising than same polarity charge repulsions (e.g. O5-7 or O13-15 > O2-4, O10-12, O1-3 or O9-11, ΔT_m = 9 to 48; O2-3 or O9-12 > O6-7, O13-16, O5-8, or O14-15, ΔT_m = 1 to 26). This suggests that ***for off-targets, intra-molecular attraction decreases inter-molecular repulsion and increases CC stability***. Similarly, when there is both intra- and inter-molecular repulsion (i.e. ++ or --) the T_m is decreased. We speculate that alternating charges at ***e*** or ***g*** positions promotes intra-molecular attraction and leads to decreased ***g-e'***$^{+1}$ inter-molecular repulsion. These effects increase stability for these intra-pair off-targets. This pattern is observed throughout the off-target sets. Alternating intra-molecular charges are therefore to be disfavoured for both desired states and off-targets when designing heterospecific sets, and function in an opposite but analogous way (Figure 4).
- ***For homodimeric off-targets (Table S2), negative inter-molecular charge repulsions towards the N-terminus (i.e. E-E g-e'***$^{+1}$ ***pairs) and positive charge repulsions (i.e. K-K g-e'***$^{+1}$ ***pairs) at the C-***

terminus generally generated increased stability. Reversing this pattern generally destabilises the CCs. In general, it is more stabilising for the CC to have a negative N-terminus than to have a positive C-terminus (Q1-1 > Q2-2 (ΔT_m = 12), O9-9 > O1-1 (ΔT_m = 8), O3-3 > O11-11 (ΔT_m = 9), O2-2 > O10-10 (ΔT_m = 5), Q8-8 > Q7-7 (ΔT_m = 16)) [37]. The effect of placing a two positive repulsive pair at the C-terminus is not clear. This pattern of negative charge at the N-terminus and positive charge at the C-terminus adding stability generally holds for inter-family off-targets (e.g. O12-14 *vs*. O3-5 or O4-6) and is most pronounced when the electrostatics are fully repulsive (e.g. O5-15 *vs*. O6-16).

On the basis of these findings, in future solid blocks of three or more E/K residues (i.e. at three consecutive ***e*** or ***g*** positions) should be included in peptide library designs since they will assist in stabilising desired pairs while concomitantly destabilising off-targets, leading to a favourable increase in ΔT_m (desired – off-target) (Figure 4). In addition, introducing E at the N-terminus and (less so) K at the C-terminus will further aid stability. Alternating charge repulsions on the same helix should be avoided since they promote intra-helical electrostatics; this will have the effect of both reducing inter-molecular repulsion for off-targets while also reducing the inter-molecular electrostatic attractions in the desired states. Although these observations present general trends, it is difficult to predict the magnitude of the effects in building sequence-specific context into stability prediction models. Nonetheless, creating libraries that conform to these 'charge block' rules in the first instance means they no longer need to be explicitly searched for. Rather, by defining permitted ***e*** and ***g*** charge block arrangements (i.e. EEEE/KKKK/KEEE/EKKK/EEEK/KKKE) with the same core arrangement as previously specified (i.e. (6*7/2)*6cores = 126 member library) we are able to screen an interactome of 8001 potential CCs. This resulted in the identification of twelve sets of decuples (10 CCs) based on the same cut-off parameters as used in Table 2.

Improvements in the speed and flexibility of the software mean that many new avenues of *in silico* screening are now possible, with key patterns in the interaction profiles visible from an observational level (Table 1). The added role of sequence context is of interest as it can further improve the prediction of heterospecific peptides by added to an increased energy gap between desired and non-desired CCs. In order to further analyse and predict interaction stability based on these patterns a larger training set would be required.

Conclusion

We have increased the capacity of our predictive algorithm to identify a set of sixteen new peptides capable of forming eight heterospecific CC pairs. Of these, seven have been demonstrated to function

as predicted. To our knowledge this is the largest heterospecific set of designed peptides created to date. In expanding the predicted heterospecific set from eight peptides [22] to sixteen we have:

i) By necessity, increased both the speed and utility of the algorithm. Although we have stopped at Octuples, by continuing with our Python-based approach we have expanded our predicted heterospecific CC set up to duodecuples (12 CCs) using the current library.

ii) Implemented the removal of peptides predicted to form antiparallel CCs (i.e. those that can adopt fully-complementary ***e-e'*** or ***g-g'*** electrostatic pairs by the algorithm).

iii) Robustly demonstrated both the need for such software and its utility in directing against the expanding the number of lower-energy off-targets, in this case from 32 to 128. The new Python-based algorithm predicts that we can further expand the number of peptides to at least 24. This would generate 12 heterospecific CCs, with 288 CC off-targets, using the same two core (Ile/Asn) and electrostatic (Glu/Lys) residue options.

iv) Used our data set of >170 CCs to identify electrostatic 'charge-blocks' (Figure 4). These aid the *de novo* design of specificity by serving to increase the stability of desired pairs, while concomitantly decreasing off-target stability. Designing charge-blocks into future CC based systems will assist in ensuring that designed peptide sets achieve their desired heterospecificity. Incorporating these and other emerging sequence context based rules for otherwise energetically equivalent CCs into prediction models will further ensure that off-targets are disfavoured while increasing the predicted stability of desired pairs.

v) Lastly, we believe that the heterospecific peptide sequences generated and the tools used to identify them will also be of use to the synthetic biology community. As more data becomes available, we will expand the size of both training and test-sets to further increase CC prediction accuracy. The software and peptides derived from the study, as well as the approach more widely, has the potential to be applied in a variety of downstream applications that include hydrogels, increased complexity nanocages, PPI inhibitors, and as peptide-tags for uses as molecular probes [3-5].

Our aim to derive a heterospecific interactome using 16 peptides was partially achieved; with 7 of the 8 CCs shown to be heterospecific. However, observations from the expanded dataset have given rise to a significant increase in the accuracy of CC prediction. Incorporating emerging rules into qCIPA selection to screen and select large heterospecific peptide sets represents a significant advance towards designing interactomes that are more likely to be exquisitely specific. In future it may be possible to further improve the accuracy of specificity prediction by taking into account additional coupling energies or by accounting for context dependence of additional residue-pair interactions [38 39]. We believe that these findings make important contributions to the question of how primary sequence

governs the stability and specificity of quaternary structures, and in the derivation of peptide building blocks to modulate PPIs as well as tools for the synthetic biology community.

Supporting Information

Additional Information is given in the Supporting Information File. This includes details on the software (including web links), peptide sequences, and additoonal information on the parameters used to identify Octuples. This material is available free of charge via the Internet at http://pubs.acs.org.

Acknowledgements

JMM is grateful to Cancer Research UK for a Career Establishment Award (C29788 / A11738) and to the Engineering and Physical Sciences Research Council for an Overseas Travel Grant (EP/M001873/2).

Conflict of Interest

The authors declare that they have no conflicts of interest with the contents of this article.

Author Contributions

J.M. suggested and supervised the work. J.M. and R.C. designed and synthesised the peptides. R.C. acquired the CD data. R.C. and J.M. analysed the CD data. R.C. and A.L. wrote and implemented the software. A.P. acquired and analysed the SEC data. J.M. wrote the paper with input from all authors.

REFERENCES

[1] Reinke, A. W., Grant, R. A., and Keating, A. E. (2010) A synthetic coiled-coil interactome provides heterospecific modules for molecular engineering, *J. Am. Chem. Soc. 132*, 6025-6031.

[2] Boyle, A. L., Bromley, E. H., Bartlett, G. J., Sessions, R. B., Sharp, T. H., Williams, C. L., Curmi, P. M., Forde, N. R., Linke, H., and Woolfson, D. N. (2012) Squaring the circle in peptide assembly: from fibers to discrete nanostructures by de novo design, *J. Am. Chem. Soc. 134*, 15457-15467.

[3] Shlizerman, C., Atanassov, A., Berkovich, I., Ashkenasy, G., and Ashkenasy, N. (2010) De Novo Designed Coiled-Coil Proteins with Variable Conformations as Components of Molecular Electronic Devices, *Journal of the American Chemical Society 132*, 5070-5076.

[4] Gradisar, H., Bozic, S., Doles, T., Vengust, D., Hafner-Bratkovic, I., Mertelj, A., Webb, B., Sali, A., Klavzar, S., and Jerala, R. (2013) Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments, *Nat Chem Biol 9*, 362-366.

[5] Fairman, R., and Akerfeldt, K. S. (2005) Peptides as novel smart materials, *Curr Opin Struc Biol 15*, 453-463.

[6] Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, J. D., and Voelz, V. A. (2007) The protein folding problem: when will it be solved?, *Curr Opin Struc Biol 17*, 342-346.

[7] Mason, J. M., and Arndt, K. M. (2004) Coiled coil domains: stability, specificity, and biological implications, *Chembiochem 5*, 170-176.

[8] Woolfson, D. N. (2005) The design of coiled-coil structures and assemblies, *Adv. Protein Chem. 70*, 79-112.

[9] Krylov, D., Barchi, J., and Vinson, C. (1998) Inter-helical interactions in the leucine zipper coiled coil dimer: pH and salt dependence of coupling energy between charged amino acids, *J. Mol. Biol. 279*, 959-972.

[10] Acharya, A., Ruvinov, S. B., Gal, J., Moll, J. R., and Vinson, C. (2002) A heterodimerizing leucine zipper coiled coil system for examining the specificity of a position interactions: amino acids I, V, L, N, A, and K, *Biochemistry 41*, 14122-14131.

[11] Acharya, A., Rishi, V., and Vinson, C. (2006) Stability of 100 homo and heterotypic coiled-coil a-a ' pairs for ten amino acids (A, L, I, V, N, K, S, T, E, and R), *Biochemistry 45*, 11324-11332.

[12] Williams, R. W., Chang, A., Juretic, D., and Loughran, S. (1987) Secondary structure predictions and medium range interactions, *Biochim Biophys Acta 916*, 200-204.

[13] Newman, J. R., and Keating, A. E. (2003) Comprehensive identification of human bZIP interactions with coiled-coil arrays, *Science 300*, 2097-2101.

[14] Mason, J. M., Schmitz, M. A., Muller, K. M., and Arndt, K. M. (2006) Semirational design of Jun-Fos coiled coils with increased affinity: Universal implications for leucine zipper prediction and design, *Proc. Natl. Acad. Sci. USA 103*, 8989-8994.

[15] Hagemann, U. B., Mason, J. M., Muller, K. M., and Arndt, K. M. (2008) Selectional and mutational scope of peptides sequestering the Jun-Fos coiled-coil domain, *J. Mol. Biol. 381*, 73-88.

[16] Fong, J. H., Keating, A. E., and Singh, M. (2004) Predicting specificity in bZIP coiled-coil protein interactions, *Genome Biol 5*, R11.

[17] Thompson, K. E., Bashor, C. J., Lim, W. A., and Keating, A. E. (2012) SYNZIP protein interaction toolbox: in vitro and in vivo specifications of heterospecific coiled-coil interaction domains, *Acs Synth Biol 1*, 118-129.

[18] Grigoryan, G., Reinke, A. W., and Keating, A. E. (2009) Design of protein-interaction specificity gives selective bZIP-binding peptides, *Nature 458*, 859-864.

[19] Potapov, V., Kaplan, J. B., and Keating, A. E. (2015) Data-driven prediction and design of bZIP coiled-coil interactions, *PLoS Comput Biol 11*, e1004046.

[20] Negron, C., and Keating, A. E. (2014) A Set of Computationally Designed Orthogonal Antiparallel Homodimers that Expands the Synthetic Coiled-Coil Toolkit, *Journal of the American Chemical Society 136*, 16544-16556.

[21] Wood, C. W., Bruning, M., Ibarra, A. A., Bartlett, G. J., Thomson, A. R., Sessions, R. B., Brady, R. L., and Woolfson, D. N. (2014) CCBuilder: an interactive web-based tool for building, designing and assessing coiled-coil protein assemblies, *Bioinformatics 30*, 3029-3035.

[22] Crooks, R. O., Baxter, D., Panek, A. S., Lubben, A. T., and Mason, J. M. (2016) Deriving Heterospecific Self-Assembling Protein-Protein Interactions Using a Computational Interactome Screen, *J. Mol. Biol. 428*, 385-398.

[23] Crooks, R. O., Rao, T., and Mason, J. M. (2011) Truncation, Randomization, and Selection: Generation of a Reduced Length c-Jun Antagonist That Retains High Interaction Stability, *J. Biol. Chem. 286*, 29470-29479.

[24] Harbury, P. B., Zhang, T., Kim, P. S., and Alber, T. (1993) A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants, *Science 262*, 1401-1407.

[25] Oakley, M. G., and Kim, P. S. (1998) A buried polar interaction can direct the relative orientation of helices in a coiled coil, *Biochemistry 37*, 12603-12610.

[26] Gurnon, D. G., Whitaker, J. A., and Oakley, M. G. (2003) Design and characterization of a homodimeric antiparallel coiled coil, *J. Am. Chem. Soc. 125*, 7518-7519.

[27] O'Shea, E. K., Klemm, J. D., Kim, P. S., and Alber, T. (1991) X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil, *Science 254*, 539-544.

[28] Gradisar, H., and Jerala, R. (2011) De novo design of orthogonal peptide pairs forming parallel coiled-coil heterodimers, *J. Pept. Sci. 17*, 100-106.

[29] McClain, D. L., Woods, H. L., and Oakley, M. G. (2001) Design and characterization of a heterodimeric coiled coil that forms exclusively with an antiparallel relative helix orientation, *J. Am. Chem. Soc. 123*, 3151-3152.

[30] Oakley, M. G., and Hollenbeck, J. J. (2001) The design of antiparallel coiled coils, *Curr. Opin. Struct. Biol. 11*, 450-457.

[31] Fields, G. B., and Noble, R. L. (1990) Solid-Phase Peptide-Synthesis Utilizing 9-Fluorenylmethoxycarbonyl Amino-Acids, *Int J Pept Prot Res 35*, 161-214.

[32] Elwell, M. L., and Schellman, J. A. (1977) Stability of phage T4 lysozymes. I. Native properties and thermal stability of wild type and two mutant lysozymes, *Biochim Biophys Acta 494*, 367-383.

[33] Mason, J. M., Hagemann, U. B., and Arndt, K. M. (2007) Improved stability of the Jun-Fos Activator Protein-1 coiled coil motif: A stopped-flow circular dichroism kinetic analysis, *J. Biol. Chem. 282*, 23015-23024.

[34] O'Shea, E. K., Lumb, K. J., and Kim, P. S. (1993) Peptide 'Velcro': design of a heterodimeric coiled coil, *Curr. Biol. 3*, 658-667.

[35] Woolfson, D. N., and Alber, T. (1995) Predicting oligomerization states of coiled coils, *Protein Sci 4*, 1596-1607.

[36] Cochran, D. A. E., and Doig, A. J. (2001) Effect of the N2 residue on the stability of the alpha-helix for all 20 amino acids, *Protein Sci. 10*, 1305-1311.

[37] Doig, A. J., and Baldwin, R. L. (1995) N- and C-capping preferences for all 20 amino acids in alpha-helical peptides, *Protein Sci 4*, 1325-1336.

[38] Steinkruger, J. D., Woolfson, D. N., and Gellman, S. H. (2010) Side-chain pairing preferences in the parallel coiled-coil dimer motif: insight on ion pairing between core and flanking sites, *J. Am. Chem. Soc. 132*, 7586-7588.

[39] Steinkruger, J. D., Bartlett, G. J., Woolfson, D. N., and Gellman, S. H. (2012) Strong contributions from vertical triads to helix-partner preferences in parallel coiled coils, *J. Am. Chem. Soc. 134*, 15652-15655.

FIGURE LEGENDS

Figure 1: Software to computationally guide derivation of specific coiled coils. Firstly '*Generate Library'* was used to create a complete list of the peptide library (in this case a 4,096 member peptide library was reduced to 1,536). Next the '*bCIPA Interactome Screen'* was used to predict the T_m of each potential CC within the 1,180,416 pairwise CC interactome. Next *'Find Pairs'* was used to identify sets of peptides that, according to the criteria input by the user, are predicted to be heterospecific when combined. In this case 510 such sets were identified. Next '*Find Quadruples'* was used to combine pairs of CCs to identify groups of four CCs that are predicted to be heterospecific when combined. Here, 15,171 sets were identified. Finally, *'Find Octuples'* combined sets of Quadruples to identify 27,501 sets of 16 peptides that are predicted to be heterospecific when mixed. The sixteen-peptide set used in this study is shown, with additional capping sequences underlined. Shown on the right are the bCIPA predicted and measured thermal melting values for all 136 possible pairs within the selected sixteen-peptide interactome. All peptides are distinct from those in our previous Quadruple set[22]. For a full description of the software see the Supporting Information.

Figure 2: Thermal stability of peptide pairs measured by using temperature dependence of the CD signal at 222 nm. All 136 peptide pairs are shown, with heterospecific pairs colour coded according to the key. The data demonstrates that with the exception of 15-16, all desired peptides display T_m values that are higher than any measured off-target pair.

Figure 3: Bar chart and heat maps displaying T_m values predicted by bCIPA and qCIPA as well as those experimentally measured. The values have been grouped according to the core and electrostatic arrangements. These are desired pairs (2II 2NN core, all attractive electrostatics), homodimeric off-targets (2II 2NN core, all repulsive electrostatics, Intra-pair off-targets (4NI core, 4 attractive / 4 repulsive electrostatics), Inter-pair off-target 1 (1II 1NN 2NI core, 4 attractive / 4 repulsive electrostatics), Inter-pair off-target 2 (2II 2NN core, 4 attractive / 4 repulsive electrostatics), Inter-pair off-target 3 (4NI core, all attractive electrostatics), and Inter-pair off-targets 4 (4NI core, all repulsive electrostatics). qCIPA was derived by using the 22 CCs from our previous work in this area[22] as a training set. Instead of Core and electrostatic weightings used by bCIPA, qCIPA uses individual weightings for II/NN/IN (core) and EE/KK/EK (electrostatic) arrangements, and results in an improved fit to the training set. See also Tables s1-7 for a comprehensive list of core/electrostatic combinations that fall into one of the 7 categories described above.

Figure 4: Effect of electrostatic charge blocks on the predicted T_m. For desired pairs blocks of 3-4 consecutive same charge residues at either ***e*** or ***g*** positions on each peptide led to increased stability over that purely based on the sum of the core and electrostatic components. We speculate that these 'charge blocks' have two benefits; ***i)*** they increase stability for these desired pairs by promoting inter-molecular attraction (e.g. A *vs*. B) ***ii)*** in a similar but opposite manner, charge-blocks decrease the

stability of off-target CCs by promoting inter-molecular repulsion (e.g. C *vs.* D). The net effect is therefore that the ΔT_m(desired – off-targets) is increased when 'charge blocks' are introduced (i.e. A-D > B-C).

	Quadruples		Octuples		Octuples with arrangement	Predicted Tm range bCIPA	Predicted Tm range qCIPA	Measured Tm range
	Core	**Electrostatics**	**Core**	**Electrostatics**				
Desired Pairs	2II, 2NN ΔG = -23.2	8EK/KE ΔG = -9.6	2II 2NN ΔG = -23.2	8EK/KE ΔG = -9.6	8	73 °C ΔG = -32.8	62 °C	52 °C to 75 °C
Homodimeric Off-targets	2II, 2NN ΔΔG = 0	8KK/EE ΔΔG = +7.2 to +12.8	2II 2NN ΔΔG = 0	8KK/EE ΔΔG = +7.2 to +12.8	16	-8 to 1 °C ΔΔG = = +7.2 to +12.8	17 to 30 °C	-2 °C to 34 °C
Intra-pair off-targets	4 NI ΔΔG = +21.2	4KE, 4EE/KK ΔΔG = +3.6 to +6.4	4NI ΔΔG = +21.2	4KE, 4EE/KK ΔΔG = +3.6 to +6.4	16	9 to 18 °C ΔΔG = +24.8 to +27.6	20 to 33 °C	3 °C to 51 °C
Inter-pair off-targets	1II, 1NN, 2NI ΔΔG = +10.6	4KE, 4EE/KK ΔΔG = +3.6 to +6.4	1II, 1NN, 2NI ΔΔG = +10.6	4KE, 4EE/KK ΔΔG = +3.6 to +6.4	64	20 to 28 °C ΔΔG = +14.2 to +17	28 to 41 °C	-3 °C to 60 °C
			2II 2NN (new) ΔΔG = 0	*4KE, 4EE/KK (new)* ΔΔG = +3.6 to +6.4	16	30 to 39 °C ΔΔG = +3.6 to +6.4	36 to 39 °C	12 to 60 °C
			4NI (new) ΔΔG = +21.2	*8KE (new)* ΔΔG = 0	8	52 °C ΔΔG = +21.2	46 °C	10 to 42 °C
			4NI (new) ΔΔG = +21.2	*8KK/EE (new)* ΔΔG = +7.2 to +12.8	8	-29 to -21 °C ΔΔG = +28.4 to +34	1 to 14 °C	1 to 32 °C

Table 1: Energetic arrangements found in Quadruples are expanded in Octuples sets. The doubling in the number of desired heterospecific coiled coil pairs leads to a loss in stringency and therefore specificity of interaction as the number of off-targets is quadrupled. The three new Inter-pair off-target combinations are 2II 2NN core with four attractive / 4 repulsive electrostatics, and all attractive /all repulsive ES with 4NI core mismatches, resulting in bCIPA estimating the Tm to be 52 °C or -21/-29 °C respectively. Shown are the contributions to folding from core and electrostatic interactions, as well as their sum. All free energies are shown in kcal/mol and are based on free energy scores derived from a double mutant analysis.

Tuple	Numbers of Sets in PHP system (parameters used)	Numbers of Sets in Python system (parameters used)	Numbers of Sets in Python system (parameters used)
1	**36** (70/10/20/50)	**36** (70/10/52/21)	**36** (66/14/50/23)
2	**42**	**510**	**492**
3	N/A	**3708**	**3264**
4	**72** (70/10/52/21)	**15171**	**11067**
5	N/A	**36204**	**18180**
6	N/A	**51450**	**11430**
7	N/A	**45456**	**0**
8	**36**	**27501**	**0**
9	N/A	**11904**	**0**
10	N/A	**3552**	**0**
11	N/A	**648**	**0**
12	**0**	**54**	**0**

Table 2: The number of unique sets identified using the previous php-based approach vs. a python based approach. Both data sets implement the exclusion of predicted high affinity antiparallel dimers and specify minimum/maximum of 2 Asn residue to confer maximal core specify upon selected sets. For php, the approach of combining sets (e.g. 2→4→8) was necessary to reduce the computing time required to generate results. In contrast, python adds one coiled coil at a time, resulting in much higher numbers of predicted heterospecific sets. By using the settings required at the quadruples stage from the beginning (column 2 *vs.* column 1) we founds that once again more coiled coils were permitted to progress through each round to arrive at duodecuples, further demonstrating the benefit of early redundancy in the system. Further increasing the stringency (minimum ΔT_m) resulted in no heterospecific sets beyond sextuples (column 3). Column 2 took python approximately 4 days to run on a single PC. Shown in parentheses are the minimum desired T_m / minimum homodimer T_m / minimum off-target T_m / minimum ΔT_m, requested for each round respectively.

Figure 1

Generate Peptide Library Sequences

1536 peptides

↓

bCIPA Interactome Screen

1,180,416 PPIs

↓

Find Coiled-coil Pairs (2)

510 sets identified

↓

Find Coiled-coil Quadruples (4)

15,171 sets identified

↓

Find Coiled-coil Octuples (8)

27,501 sets identified

↓

```
Oct1:    ASKNAALKAENAALEYEIAALEAEIAALEGAP
Oct2:    ASENAALEAKNAALKYKIAALKAKIAALKGAP
Oct3:    ASEIAALEAKIAALKYENAALEAENAALEGAP
Oct4:    ASKIAALKAEIAALEYKNAALKAKNAALKGAP
Oct5:    ASENAALEAEIAALEYKNAALKAEIAALEGAP
Oct6:    ASKNAALKAKIAALKYENAALEAKIAALKGAP
Oct7:    ASKIAALKAKNAALKYKIAALKAENAALEGAP
Oct8:    ASEIAALEAENAALEYEIAALEAKNAALKGAP
Oct9:    ASENAALEAKNAALKYEIAALEAEIAALEGAP
Oct10:   ASKNAALKAENAALEYKIAALKAKIAALKGAP
Oct11:   ASKIAALKAEIAALEYENAALEAENAALEGAP
Oct12:   ASEIAALEAKIAALKYKNAALKAKNAALKGAP
Oct13:   ASKNAALKAKIAALKYKNAALKAEIAALEGAP
Oct14:   ASENAALEAEIAALEYENAALEAKIAALKGAP
Oct15:   ASEIAALEAENAALEYKIAALKAENAALEGAP
Oct16:   ASKIAALKAKNAALKYEIAALEAKNAALKGAP
```

Eight predicted Heterospecific Coiled coils: 'Octuples' →

Predicted by bCIPA

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	1	73	18	13	28	24	24	28	39	34	-21	52	24	28	28	24	1
2		-8	13	9	24	20	20	24	34	30	52	-29	20	24	24	20	2
3			1	73	28	24	24	28	-21	52	39	34	24	28	28	24	3
4				-8	24	20	20	24	52	-29	34	30	20	24	24	20	4
5					1	73	13	18	28	24	28	24	34	39	-21	52	5
6						-8	9	13	24	20	24	20	30	34	52	-29	6
7							-8	73	24	20	24	20	-29	52	34	30	7
8								1	28	24	28	24	52	-21	39	34	8
9									1	73	18	13	24	28	28	24	9
10										-8	13	9	20	24	24	20	10
11											1	73	24	28	28	24	11
12												-8	20	24	24	20	12
13													-8	73	13	9	13
14														1	18	13	14
15															1	73	15
16																-8	16

↓ **Measure**

Observed

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	-2	72	3	20	21	23	34	-3	12	42	1	36	43	23	8	14	1
2		34	27	21	30	30	15	51	56	46	38	17	13	54	42	18	2
3			13	68	30	33	18	20	18	10	16	60	22	30	6	21	3
4				21	36	38	16	22	28	32	50	36	28	36	40	15	4
5					17	71	35	26	11	48	15	29	39	46	36	21	5
6						23	19	48	40	27	23	30	44	55	31	25	6
7							-1	68	17	29	21	15	18	56	40	15	7
8								2	22	29	5	48	42	17	28	28	8
9									6	69	5	42	32	18	20	18	9
10										29	31	26	26	31	26	28	10
11											4	73	26	17	13	4	11
12												15	20	51	26	21	12
13													21	75	51	16	13
14														24	20	34	14
15															12	52	15
16																0	16

Figure 2

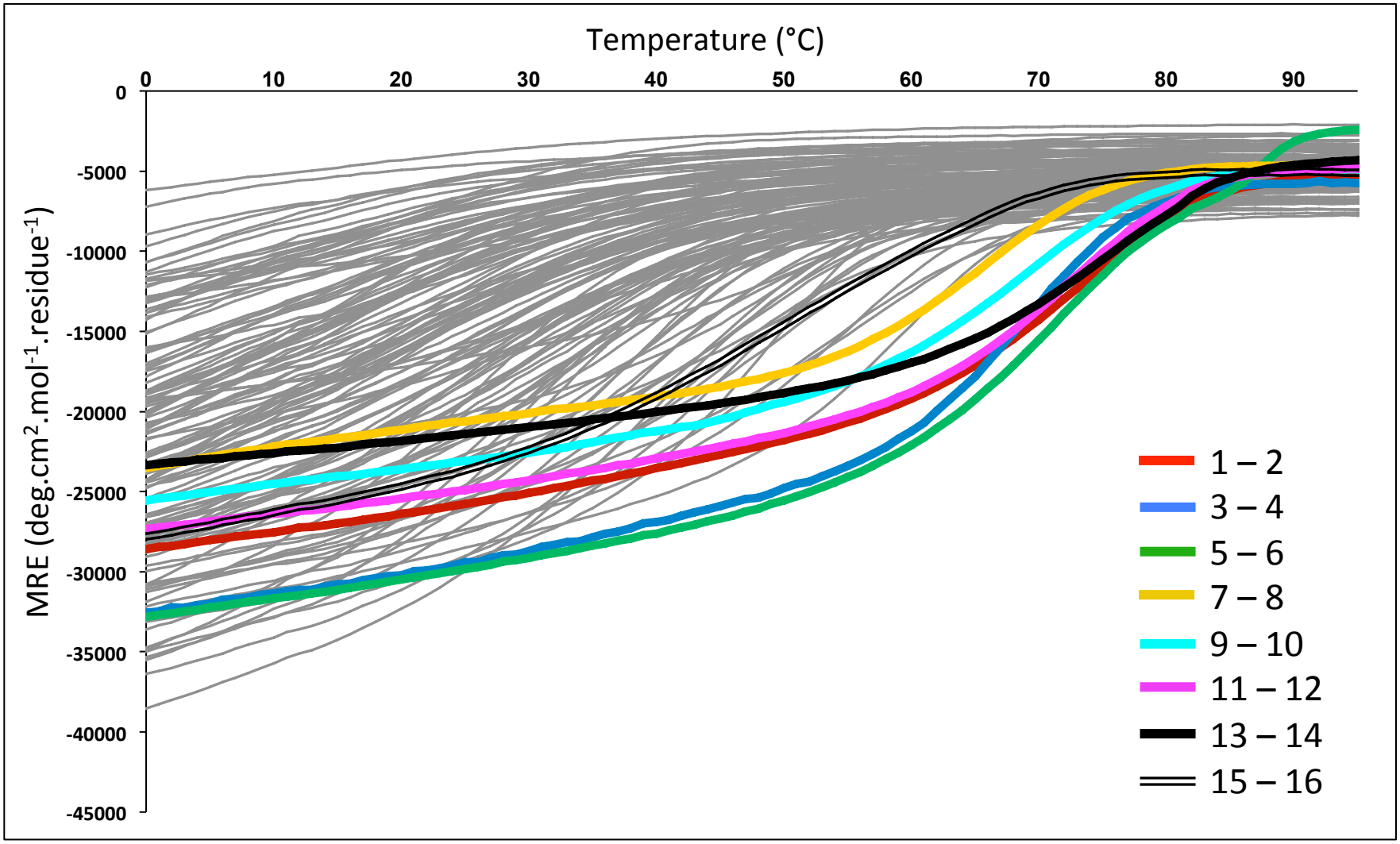
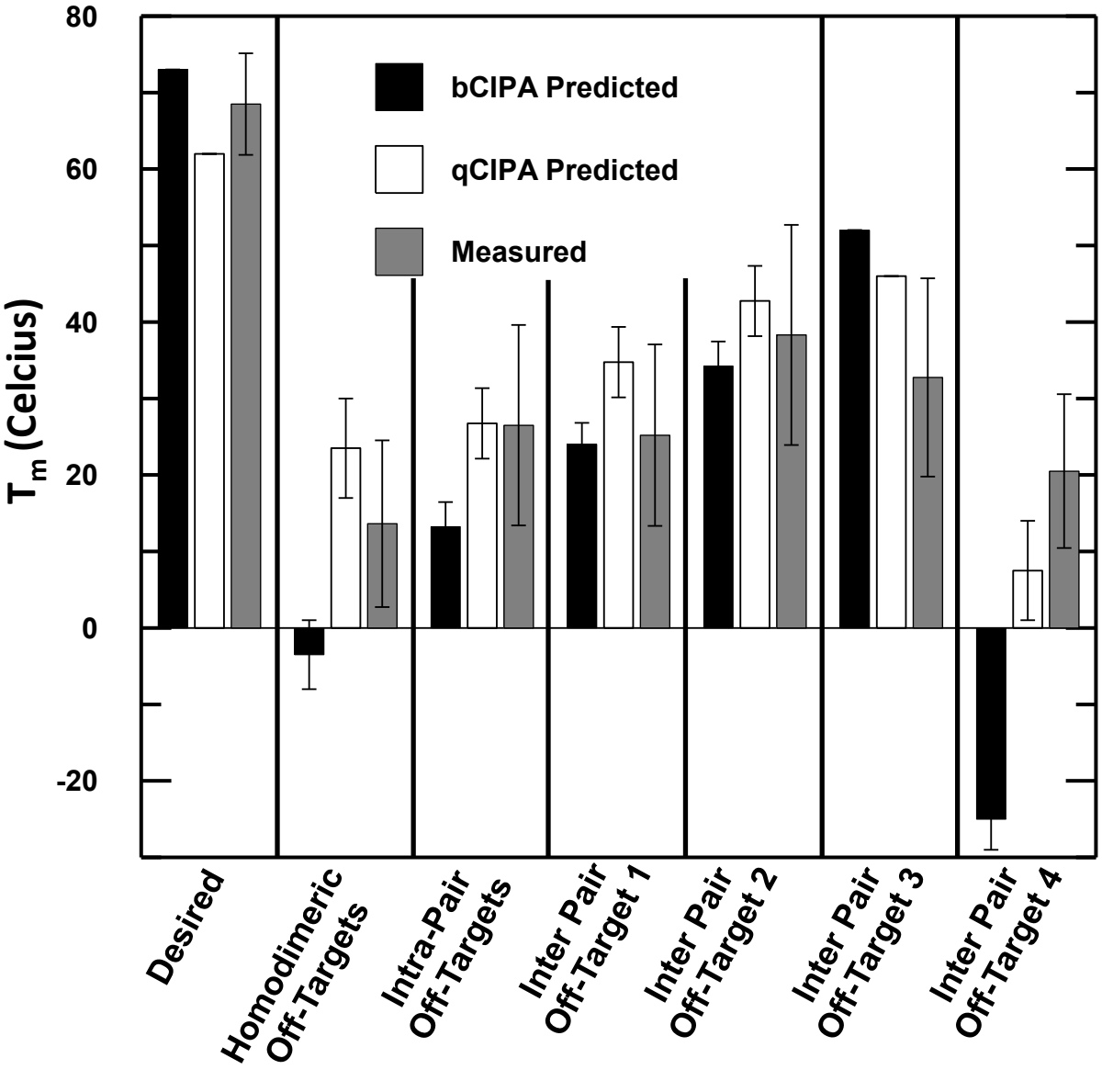
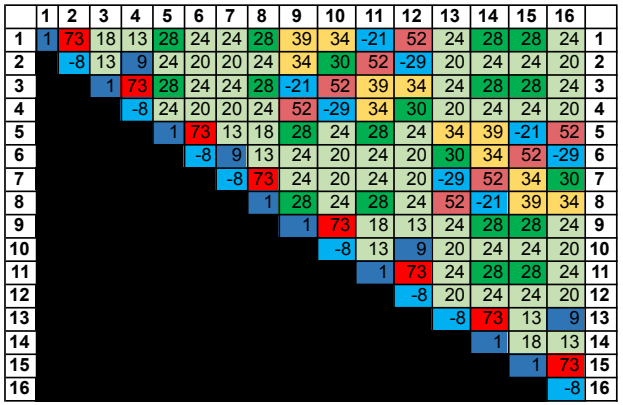
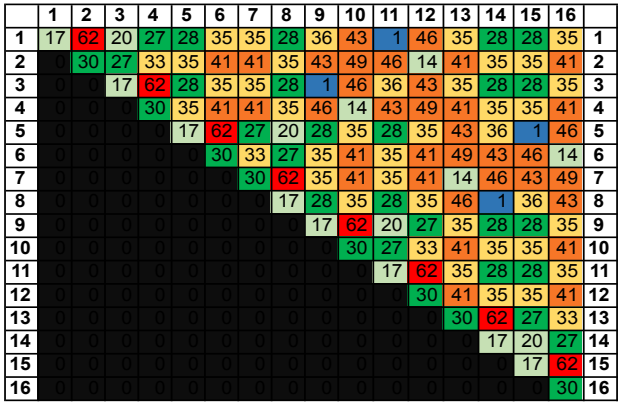

Temperature (°C)
0
10
20
30
40
50
60
70
80
90
0
-5000
-10000
-15000
-20000
-25000
-30000
-35000
-40000
-45000
MRE (deg.cm2.mol-1.residue-1)
1 – 2
3 – 4
5 – 6
7 – 8
9 – 10
11 – 12
13 – 14
15 – 16

Figure 3


80
60
40
20
0
-20
T_m (Celcius)
bCIPA Predicted
qCIPA Predicted
Measured
Desired
Homodimeric Off-Targets
Intra-Pair Off-Targets
Inter Pair Off-Target 1
Inter Pair Off-Target 2
Inter Pair Off-Target 3
Inter Pair Off-Target 4


bCIPA Predicted


qCIPA Predicted

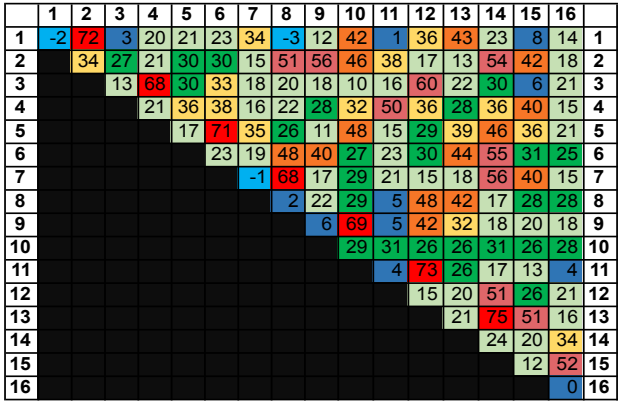
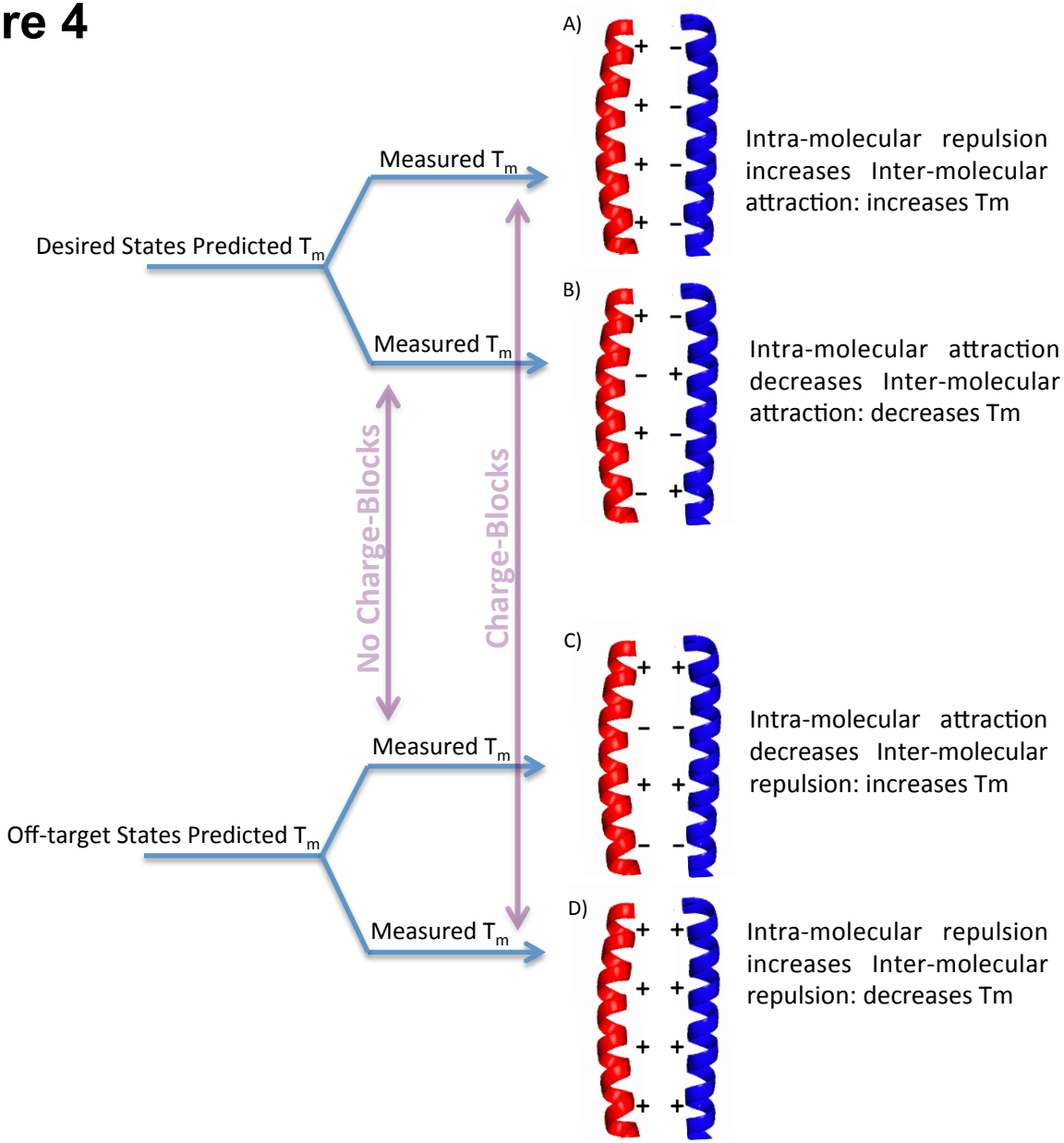

Measured

Figure 4


Desired States Predicted T_m
Measured T_m
Measured T_m
No Charge-Blocks
Charge-Blocks
Off-target States Predicted T_m
Measured T_m
Measured T_m
A)
Intra-molecular repulsion increases Inter-molecular attraction: increases Tm
B)
Intra-molecular attraction decreases Inter-molecular attraction: decreases Tm
C)
Intra-molecular attraction decreases Inter-molecular repulsion: increases Tm
D)
Intra-molecular repulsion increases Inter-molecular repulsion: decreases Tm